Homework 1: 40pts

Assigned: September 15, 2021

Due: September 27, 2021, 8:30am

1. (**5pts**) Let the random sample and let .
2. (1pt) What is , the distribution of (in terms of and )?

The distribution of X is Normal, with mean = \mu\_Y and variance \sigma\_Y^2

1. (2pts) Let be a random sample of , where is the distribution you just specified in part (a). Let . What is , the distribution of ?

The distribution of W is also Normal

(2pts) Let . What is , the distribution of ?T distribution with n-1 degrees of freedom

1. (**3pts**) Suppose and are the respective PDF and CDF of a random variable .
2. (1pt) Describe in words.

F\_y(3) would be the probability of obtaining a 3 on a single random draw of the random variable Y.

1. (1pt) Describe in words.

F\_y(3) would be the probability of obtaining a number 3 or less on a single random draw

1. (1pt) If is the th quantile of , what is ?

F\_y(z\_p) is the cumulative probability of obtaining a number less than or equal to the number corresponding to the *p*th quantile of Y

1. (**5pts**) Give the values of the following.
2. (2pts) Find , , , , and to 3 decimals.

1.960, 2.228, 2.009, 1.984, 1.965

1. (3pt) Find (29 degrees of freedom is like a sample of size ). Comment on why using the appropriate sampling distribution matters.

2.045, the shape of the distribution changes across different distributions, and even with in the t distribution, the shape differs with a change in the degrees of freedom. Changing the shape of a distribution changes the values computed for quantiles and probability.

1. (**6pts**) Suppose the number of apneic episodes per hour for individuals being treated with an experimental drug for sleep apnea is .
2. (1pts) What is the probability that a person chosen at random from this population will have more than 3 episodes per hour?

0.0483

1. (1pt) What is the probability that a person chosen at random will have between 1 and 3 episodes per hour?

0.4517

1. (1pt) What is the 11th percentile of apneic episodes per hour for this distribution?

-.0262

1. (3pts) Based on your answer in (c), what is an issue with assuming that apneic episodes per hour is normally distributed?

An issue with this is that it may not be normally distributed, because a negative number for the 11th percentile should not be possible logically when considering it represents the number of apneic episodes per hour.

(**21pts**) FINAL PROJECT

The file “NHANES702.csv” (located in Resources → Final Project on Sakai) contains data for the Final Project described in detail during class. You may review this in the recording on 9/1/2021, starting around the 4:25 mark.

1. (**4pts**) CONSORT numbers
2. (1pt) How many participants were missing CKD status and excluded?

47

1. (1pts) Of those with CKD data, how many were missing serum K and excluded?

3

1. (2pts) Based on your answers, how many are eligible for further analyses, according to the inclusion/exclusion criteria in the SAP?

300

1. (**9pts**) Graphical Displays
2. (5pts) Which of the following variables are approximately normally distributed and which are not? Give two plots for each variable and describe why or why not: age, BMI, serum K, diet K, eGFR.

Chart, box and whisker chart

Description automatically generated

Age is not approximately normally distributed because the shape of the histogram does not follow a normal distribution.

Chart, histogram, box and whisker chart

Description automatically generated

BMI is also not normally distributed, it is right-skewed.

Chart, histogram, box and whisker chart

Description automatically generated

The distribution of serumK is approximately normal due to these plots.

Chart, box and whisker chart

Description automatically generated

The distribution of DietK1000 is approximately normal.

Chart, box and whisker chart

Description automatically generated

The distribution of eGFR is also approximately normal, though it does show a bit of a left skew in the

1. (4pts) For the variable eGFR, is approximately 68% of the data within one standard deviation of the mean? Is approximately 95% of the data within two standard deviations of the mean? Justify your answers.

The lower 16th percentile of the data is 64.7 and the upper 16th percentile is 112.7, thus 68% of the data are between 64.7 and 112.7. The mean eGFR +- 1 SD is (65.6, 113,1) which is very close to the percentiles mentioned above.

Broadening our interval, we see that 95% of the data are between 42.6 and 131.0, and the associated mean +- 2SD is (41.7, 136.9). Both sets of these intervals are approximately equal, so we can conclude that approximately 68% of the data lie within 1 standard deviation of the mean, and that approximately 95% of the data are within 2 standard deviations of the mean.

Chart, histogram

Description automatically generated

1. (**8pts**) More exploratory analyses
2. (4pts) Within each serum K category, construct an 95% CI for the population mean eGFR. (In other words, you are constructing a CI for the population mean eGFR within two different sub populations, those with low-normal K and those with normal K.) Interpret the CIs in a way a clinician or research with no stats training would understand.

95% confidence interval for low/normal: [88.707, 96.116]. We are 95% confident that the true mean eGFR for individuals with low-normal serumK is between these two numbers.

95% confidence interval for normal: [83.046, 90.663]. We are 95% confident that the true mean eGFR for individuals with normal serumK is between these two numbers.

1. (4pts) What concerns arise when constructing confidence intervals for the population mean eGFR within diet K groups? How might you address this concern with the researcher? (Be sure to note the concern and how you handle it in the exploratory analysis section of your Final Project!)

I am concerned about the actual distribution of the data. It appears somewhat normal, but not perfectly normal. When computing the confidence interval, I assumed a normal distribution, so the calculations are based off of that distribution. The interval would be different if the distribution was not actually normal. Another potential issue is that in the categorized dietK1000 variable, there is a big imbalance in group sizes: 4,62, & 225. There are also 59 NA values. To handle this concern, we should dichotomize diet K.